What is claimed is:

1. A system for identifying language attributes through probabilistic analysis, comprising:

a storage storing a set of language classes, which each identify a language and a character set encoding, and a plurality of training documents;

an attribute modeler evaluating occurrences of one or more document properties within each training document and, for each language class, calculating a probability for the document properties set conditioned on the occurrence of the language class; and

a text modeler evaluating byte occurrences within each training document and, for each language class, calculating a probability for the byte occurrences conditioned on the occurrence of the language class.

2. A system according to Claim 1, further comprising:

a training engine calculating an overall probability for each language class by evaluating the probability for the document properties set and the probability for the byte occurrences.

3. A system according to Claim 1, further comprising:

an assignment module assigning the overall probability for each language class in accordance with the formula:

$$\arg\max_{cls} P(text \mid cls) \cdot P(props \mid cls) \cdot P(cls)$$

where *cls* is the language class, *text* is the byte occurrences set, *props* are the document properties, and $P(text \mid cls)$ is the probability for the byte occurrences, and $P(props \mid cls)$ is the probability for the document properties set.

4. A system according to Claim 1, wherein the document properties comprise at least one of top level domain, HTTP content character set encoding and language header parameters, and HTML content character set encoding and language metatags.

5. A system according to Claim 4, further comprising:

an assignment module assigning the probability for the document properties set in accordance with the formula:

$$P(tld, enc \mid cls) \cdot P(cls)$$

where *tld* is the top level domain, *enc* is the character set encoding and *cls* is the language class.

6. A system according to Claim 1, further comprising:

a counting module counting byte co-occurrences within each training document, and determining the probability for the byte occurrences based on the byte co-occurrences.

7. A system according to Claim 6, wherein the byte co-occurrences comprise a set of trigrams, further comprising:

a probability module calculating a probability of each trigram as the number of occurrences of the trigram divided by the total number of trigram occurrences in each of the training documents for each language class.

8. A system according to Claim 7, further comprising:

an assignment module assigning the probability for the byte occurrences set in accordance with the formula:

$$P(text \mid cls)$$

where *text* is the set of trigrams and *cls* is the language class.

9. A system according to Claim 1, further comprising:

a training engine performing iterative training by providing the probability for the document properties set and the probability for the byte occurrences set respectively to the evaluation of byte occurrences and assignment of the set of language classes.

10. A system according to Claim 1, further comprising:

a back off module evaluating less frequently occurring document properties by calculating a probability for each less frequently occurring document property conditioned on the occurrence of the language class.

11. A system according to Claim 1, further comprising:

a plurality of unlabeled documents; and

a classifier classifying one or more unlabeled documents by at least one language class, comprising evaluating occurrences of one or more document properties within the unlabeled document, evaluating byte occurrences within the unlabeled document, and assigning a probability for the document properties set and the byte occurrences for the unlabeled document conditioned on the occurrence of the language class.

12. A system according to Claim 11, further comprising:

a class selector selecting the at least one language class having a substantially highest probability.

13. A system according to Claim 11, further comprising:

a probability threshold; and

a pruner pruning at least one language class falling below the probability threshold.

14. A system according to Claim 1, wherein each training document comprises one of a Web page and a news message.

15. A method for identifying language attributes through probabilistic analysis, comprising:

defining a set of language classes, which each identify a language and a character set encoding, and a plurality of training documents;

evaluating occurrences of one or more document properties within each training document and, for each language class, calculating a probability for the document properties set conditioned on the occurrence of the language class; and

evaluating byte occurrences within each training document and, for each language class, calculating a probability for the byte occurrences conditioned on the occurrence of the language class.

16. A method according to Claim 15, further comprising:

calculating an overall probability for each language class by evaluating the probability for the document properties set and the probability for the byte occurrences.

17. A method according to Claim 15, further comprising:

assigning the overall probability for each language class in accordance with the formula:

$$\arg\max_{cls} P(text \mid cls) \cdot P(props \mid cls) \cdot P(cls)$$

where *cls* is the language class, *text* is the byte occurrences set, *props* are the document properties, and $P(text \mid cls)$ is the probability for the byte occurrences, and $P(props \mid cls)$ is the probability for the document properties set.

18. A method according to Claim 15, wherein the document properties comprise at least one of top level domain, HTTP content character set encoding and language header parameters, and HTML content character set encoding and language metatags.

19. A method according to Claim 18, further comprising:

assigning the probability for the document properties set in accordance with the formula:

$$P(tld, enc \mid cls) \cdot P(cls)$$

where *tld* is the top level domain, *enc* is the character set encoding and *cls* is the language class.

20. A method according to Claim 15, further comprising:

counting byte co-occurrences within each training document; and

determining the probability for the byte occurrences based on the byte co-occurrences.

21. A method according to Claim 20, wherein the byte co-occurrences comprise a set of trigrams, further comprising:

calculating a probability of each trigram as the number of occurrences of the trigram divided by the total number of trigram occurrences in each of the training documents for each language class.

22. A method according to Claim 21, further comprising:

assigning the probability for the byte occurrences set in accordance with the formula:

$$P(text \mid cls)$$

where *text* is the set of trigrams and *cls* is the language class.

23. A method according to Claim 15, further comprising:

performing iterative training by providing the probability for the document properties set and the probability for the byte occurrences set respectively to the evaluation of byte occurrences and assignment of the set of language classes.

24. A method according to Claim 15, further comprising:

evaluating less frequently occurring document properties by calculating a probability for each less frequently occurring document property conditioned on the occurrence of the language class.

25. A method according to Claim 15, further comprising:

accessing a plurality of unlabeled documents; and

classifying one or more unlabeled documents by at least one language class, comprising:

evaluating occurrences of one or more document properties within the unlabeled document;

evaluating byte occurrences within the unlabeled document; and

assigning a probability for the document properties set and the byte occurrences for the unlabeled document conditioned on the occurrence of the language class.

26. A method according to Claim 25, further comprising:

selecting the at least one language class having a substantially highest probability.

27. A method according to Claim 25, further comprising:

defining a probability threshold; and

pruning at least one language class falling below the probability threshold.

28. A method according to Claim 15, wherein each training document comprises one of a Web page and a news message.

29. A computer-readable storage medium holding code for performing the method according to Claim 15.

30. A system for identifying documents by language using probabilistic analysis of language attributes, comprising:

a set of language classes, each language class comprising a language name and a character set encoding name;

a training corpora comprising a plurality of training documents;

an attribute modeler training an attribute model by evaluating a top level domain and character set encoding associated with each training document and, for each language class, calculating a probability for each such top level domain and character set encoding conditioned on the occurrence of the each language class; and

a text modeler training a text model by evaluating co-occurrences of a plurality of bytes within each training document and, for each language class, calculating a probability for the byte co-occurrences conditioned on the occurrence of the each language class.

31. A system according to Claim 30, further comprising:

a training engine calculating an overall probability for each language class by evaluating the probability for the top level domain and character set encoding based on the attribute model and the probability for the byte occurrences based on the text model.

32. A system according to Claim 31, further comprising:

a classifier classifying one or more documents, comprising:

an attribute evaluator evaluating a top level domain and character set encoding in each document and applying the attribute model to the evaluated top level domain and character set encoding;

a text evaluator evaluating byte occurrences in each document and applying the text model to the evaluated byte occurrences; and

an assignment module assigning at least one language class based on the applications of the attribute model and the text model.

33. A method for identifying documents by language using probabilistic analysis of language attributes, comprising:

defining a set of language classes, each language class comprising a language name and a character set encoding name;

assembling a training corpora comprising a plurality of training documents;

training an attribute model by evaluating a top level domain and character set encoding associated with each training document and, for each language class, calculating a probability for each such top level domain and character set encoding conditioned on the occurrence of the each language class; and

training a text model by evaluating co-occurrences of a plurality of bytes within each training document and, for each language class, calculating a probability for the byte co-occurrences conditioned on the occurrence of the each language class.

34. A method according to Claim 33, further comprising:

calculating an overall probability for each language class by evaluating the probability for the top level domain and character set encoding based on the attribute model and the probability for the byte occurrences based on the text model.

35. A method according to Claim 34, further comprising:

classifying one or more documents, comprising:

evaluating a top level domain and character set encoding in each document and applying the attribute model to the evaluated top level domain and character set encoding;

evaluating byte occurrences in each document and applying the text model to the evaluated byte occurrences; and

assigning at least one language class based on the applications of the attribute model and the text model.

36. A computer-readable storage medium holding code for performing the method according to Claim 30.

37. An apparatus for identifying documents by language using probabilistic analysis of language attributes, comprising:

means for defining a set of language classes, each language class comprising a language name and a character set encoding name;

means for training an attribute model by assigning at least one top level domain and character set encoding pairing to at least one language class for each of a plurality of training documents and calculating a probability for each such top level domain and character set encoding pairing conditioned on the occurrence of the assigned language class; and

means for training a text model by evaluating co-occurrences of a plurality of bytes within each training document and, for each language class, calculating a probability for the byte co-occurrences conditioned on the occurrence of the language class based on the attribute model.